

LEUKEMIA CLUSTERS AND TCE WASTE SITES IN UPSTATE NEW YORK: HOW ADDING COVARIATES CHANGES THE STORY

BU-1472 -M

November, 1999

**Christina Ahrens
School of Operations Research
Cornell University**

**J.T. Gene Hwang
Department of Mathematics
Cornell University**

**Naomi Altman
Department of Biometrics
Cornell University**

**John Staudenmayer
School of Operations Research
Cornell University**

**George Casella
Department of Biometrics
Cornell University**

**Catalina Stefanescue
School of Operations Research
Cornell University**

**Malaika Eaton
Law School
Cornell University**

Keywords: *clustering, covariates, confounding, environmental epidemiology, hazardous waste sites, leukemia, trichlorethylene.*

Abstract

The New York State TCE/Leukemia dataset is often used to test new cluster detection methodologies. Examples include: Waller and Turnbull (1993), Kulldorff and Nagarwalla (1995), Waller (1996), Gangnon and Clayton (1998), Ghosh et. al (1999), and Rogerson (1999). We augment the The New York State TCE/Leukemia data with demographic covariates and find evidence of probable confounding between several of the sites and the covariates.

Leukemia Clusters and TCE Waste Sites in Upstate New York: How Adding Covariates Changes the Story

Christina Ahrens⁽¹⁾
Naomi Altman⁽²⁾
George Casella⁽²⁾
Malaika Eaton⁽³⁾
J.T. Gene Hwang⁽⁴⁾
John Staudenmayer⁽¹⁾
Catalina Stefanescu⁽¹⁾

November 28, 1999

(1) School of Operations Research, Cornell University. (2) Department of Biometrics, Cornell University. (3) Law School, Cornell University (4) Department of Mathematics, Cornell University.

Abstract: The New York State TCE / Leukemia dataset is often used to test new cluster detection methodologies. Examples include: Waller and Turnbull (1993), Kulldorff and Nagarwalla (1995), Waller (1996), Gangnon and Clayton (1998), Ghosh et al (1999), and Rogerson (1999). We augment the New York State TCE / Leukemia data with demographic covariates and find evidence of probable confounding between several of the sites and the covariates.

Keywords: Clustering; covariates; confounding; environmental epidemiology; hazardous waste sites; leukemia; trichlorethylene.

1 Introduction

Searching for clusters of disease around putative sources has long fascinated epidemiologists and biostatisticians. The famous example of John Snow mapping cholera cases in nineteenth century London and identifying certain wells as exposure foci is frequently cited in introductory epidemiology textbooks [9], [14], [16], [23]. What environmental statistician would not want to do similar work in a modern setting? While disease clusters may be associated with environmental hazards, use of clusters to infer causality is problematic. In this paper we re-examine a dataset of Waller et al [29] (and Waller's thesis [28] using data from Iwano, 1989 [12]) which was used to examine the relationship between trichloroethylene (TCE) waste sites and leukemia in upstate New York (1978–1982). Our analyses of these data illustrate one pitfall, the presence of confounding variables. The original and subsequent authors

recognized the possibility that their results were confounded, but they did not investigate further.

The paper consists of three sections. In the remainder of this section we summarize the analyses of Waller et al (1992) and subsequent re-analyses. In Section 2 we augment Waller's data with additional covariates, fit a generalized linear model, and discuss the results. The final section summarizes our results and their implications, describes the current status of the sites identified as possible disease foci, and discusses the relation between environmental statistics and public policy.

1.1 Waller et al (1992)'s Data

This subsection describes the data and the tests used in Waller et al (1992). Note that the data and several analyses also appear in the first chapter of the book *Case Studies in Biometry* (1994) [19].

TCE is an industrial solvent, suspected of contributing to leukemia incidence in exposed individuals. Direct manufacturing contact and ground-water infiltration are common exposure vectors. Although the verdict is still out on whether or not TCE is carcinogenic [15], TCE can be seen as a proxy for more dangerous industrial contamination since it is often stored with other volatile organic compounds [31].

The exposure data include the location of eleven TCE waste sites which were of environmental concern to the New York State Department of Health. These are listed in Table 1 and on the map in Figure 1.

Table 1		
Site Number	Name	County
1	Monarch Chemicals	Broome
2	IBM Endicott	Broome
3	Singer	Broome
4	Nesco	Broome
5	GE Auburn	Cayuga
6	Solvent Savers	Chenango
7	Smith Corona	Cortland
8	Victory Plaza	Tioga
9	IBM Owego	Tioga
10	Hadco	Tioga
11	Morse Chain	Tompkins

The outcome data include the number of incident leukemia cases, the population, and the location of the 790 census block groups in the counties of interest. The database of cases and addresses was built by the New York State Department of Health and contains all reported cases in the area of interest during 1978 to 1982 [12]. Since the address of some cases could only be narrowed down to the county or census tract level, some cases were fractionally allocated to several block groups by the population in those block groups [28]. The source for the population data was the 1980 U.S. census.

1.2 Application of General and Focused Tests

Waller et al (1992) apply several versions of two types of tests of spatial randomness to their data, *general* tests and *focused* tests (Besag and Newell (1991) [3]). Both types of tests address the same null hypothesis: H_0 : every person is equally likely to contract the disease independently of other cases and of the location of his or her residence.

The underlying probabilistic model considers a study region divided into I subregions with population size n_i in each subregion $i = 1, \dots, I$. For every $i = 1, \dots, I$, let C_i be a random variable representing the number of cases within subregion i .

For rare diseases such as leukemia, the null hypothesis is equivalent to: $H_0 : C_i, (i = 1, \dots, I)$ are independent Poisson random variables with $E[C_i] = \lambda n_i$. λ is the per person rate. The test types differ in their alternative hypotheses. Tests with the alternative, $H_a : \text{not } H_0$ are called *general* tests. Tests with a more specific alternative such as $H_a : E[C_i] = \lambda n_i(1 + d_i)$ (d_i is the inverse of the distance from location i to a suspected source) are called *focused* tests.

Waller et al (1992) built on Turnbull et al (1990) [25], which applied the following general tests (test without the foci) to the New York data: the GAM (geographical analysis machine) method by Openshaw et al. (1988) [21], U -statistic of Whittemore et al (1987) [33], and the cluster evaluation permutation procedure (CEPP) developed by Turnbull et al (1990) [25]. Using these tests, the evidence of clustering is weak, although there is some suggestive clustering in Cortland, Broome, and Cayuga Counties.

Waller et al (1992) then used the eleven TCE-contaminated waste sites of concern to the New York State Department of Health as the putative sources of hazard (the foci), and applied several focused tests to the data: a focused version of the method of Besag and Newell (1991) [3], a test by Waller (1992) [28], and a focused test by Stone (1988) [24].

While Besag and Newell's test does not indicate clustering around any of the eleven waste sites, Stone's procedure yields significant values at the Monarch Chemical and IBM Endicott sites in Broome County. However, neither the Stone test nor the test by Besag and Newell find statistically significant clusters when the multiplicity of tests is taken into account. In contrast, Waller's focused test (1992) gives several significant results, again showing Monarch Chemical as the focus of the most likely cluster.

1.3 Literature Review

Since 1992, these data have been reanalyzed in several published papers. Waller and Turnbull (1993) [32] used the data while discussing the effects of scale on testing for disease clustering. Kulldorff and Nagarwalla (1995) [17] applied their new general cluster detection methodology to these data and found probable clustering in Broome county. While there may have been clustering in other counties, their methodology was not designed to detect more than one cluster. Waller (1996) [30] defined the power functions for common focused tests and illustrated his results with these data. More recently, Gangnon and Clayton (1998) [5] and Ghosh et al (1999) [8] applied Bayesian cluster analysis methods to the data. Gangnon and Clayton's method found probable clustering in Broome, Cortland and Onondaga Counties. Ghosh et al.'s focused method was applied to all the sites simultaneously and found suggestive but inconclusive evidence of clustering. Rogerson (1999) [22] developed and applied chi-

squared focused and general tests to the data. He also found evidence of clustering around Monarch Chemical.

Note that none of the above studies used covariates, although most of the methods can accommodate covariate information, and all the authors noted that their analyses were potentially confounded. Of course, many of those papers were using the data to exemplify their new methodologies and epidemiological inference as not their primary concern. Recently, Waller and McMaster (1997) [31] analysed the Broome county subset of the data using counts which were externally age standardized using the National Cancer Institute's Cancer Surveillance, Epidemiology, and End Results (SEER) data [14]. They found that standardization increased the magnitude of the estimated rate around Monarch Chemical compared to an unstandardized analysis. We are aware of no other published analysis which includes any covariate information.

In this paper, we analyze these data taking covariates into consideration. Our findings, reported in Section 2.5, suggest that the effect of being near Monarch Chemical and G.E. Auburn are at least partially confounded by occupation. We find a significantly increased leukemia rate associated with living in an area with a high percentage of manufacturing employment. We also identify proximity to the Smith Corona site to be significantly associated with increased leukemia counts after controlling for other covariates. The Smith Corona site was not previously identified as a possibly dangerous site. Note that our results are still possibly confounded by unmeasured covariates.

2 Covariate Adjusted TCE Site / Leukemia Relationship

This section describes our analysis of the NYS TCE data, incorporating covariates derived from 1980 US census data. It consists of three parts. First, the additional covariate data are described. We then discuss our model-building strategy. Finally, we discuss the fitted models and conclusions based on them.

2.1 Data

We augmented Waller's 1992 data with demographic data from the 1980 census. Data at the block group level were available from two sources: Summary Tape Files 1A and 3A (STF1A and STF3A). Data from STF1A included information on the percentage of respondents in each block group of a given race, age, and house value, and whether the block group was urban or rural. STF1A had data for each of Waller et al.'s 790 block groups. STF3A recorded block group level information about education level, employment (both industry and job type), income, and source of drinking water. STF3A was missing data for 182 block groups which included all of Tompkins and Tioga counties and 21 percent of Onondaga county. Since the data did not appear to be missing at random, we did not pursue missing data strategies. Instead, we conducted two analyses: one on the complete data using only STF1A information and one on the available data for both STF1A and STF3A. The block group populations in both files matched Waller et al's data exactly.

2.2 Model Building Strategy

With leukemia rates per block group as the response, our model building strategy had three steps. First we used a Box–Cox transformation to normalize the data and chose variables for the model using all subsets linear regression. As expected, the Box–Cox procedure suggested a log transform for the leukemia rates. The variables selected (using this procedure and others described below) are listed in Table 2. Since we searched through the data to find significant variables, we need to adjust our significance levels appropriately.

Table 2: Covariates Used in the Analyses

STF1A All Subsets Selected	In Analysis	STF3A All Subsets Selected	In Analysis
% Age Over 60	Yes		Yes
Urban Indicator	No		No
% White	No		No
% Black	No		No
% House value \leq \$10K	No		No
% House value \$15-20K	No		No
% House value \$20-25K	No		No
% House value \$30-35K	No		No
		% on Public Water	Yes
		% Protection/Service Jobs	Yes
		% Other Service Jobs	No
		% Technician Jobs	Yes
		% Farming/Fishing Jobs	No
		% Precision Repair Jobs	Yes
		% Total Manufacturing Jobs	Yes
Near Monarch	Yes		Yes
Near IBM-End	No		No
Near Singer	No		No
Near Nesco	No		No
Near GE	Yes		Yes
Near Solvent	Yes		Yes
Near Smith	Yes		Yes
Near Victory	No		No
Near IBM-Owego	No		No
Near Hadco	No		No
Near Morse	Yes		No

In the second step we selected a distance function to relate the TCE sites to the cases. Partial leverage plots (Figure 2) suggested using the inverse of the distance to a site in kilometers if the block group’s centroid is within twenty kilometers of the site and zero otherwise. While such a distance function could result in infinite or poorly scaled variables, for this dataset it does not. All values are less than 3.1 km^{-1} and most are less than 1 km^{-1} . The partial leverage plots for the STF3A are in Figure 2. STF1A’s leverage plots are similar.

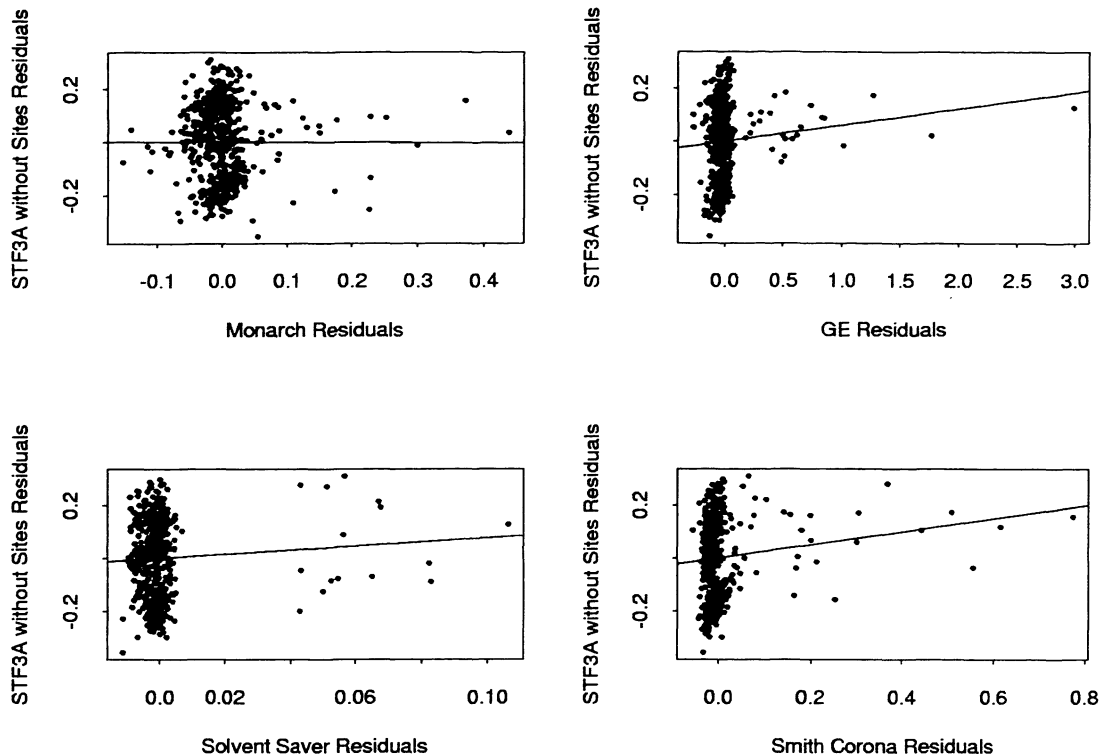


Figure 1: Partial Leverage Plots for 1/km to the TCE Sites

Once we had the set of candidate variables and the distances, our third step was to pare the list down by removing variables which were highly intercorrelated. As the map (Figure 1) suggests, several sites in the southern part of our region of interest are close together. To prevent collinearity of distance measures, we removed IBM-Endicott, Singer, Victory Plaza, IBM-Owego, Nesco, and Hadco. Since it is near the center of that region and was referred to in [28], Monarch Chemical was retained as the site from that area. The variable “Near Monarch” is a proxy for proximity to all the sites listed above.

Based on correlations between and among the sites and the other covariates, we retained the variables listed in Table 2.

2.3 The Model

We fitted generalized linear models to these data. We used a log link for the mean and a variance function which is proportional to the mean. The model corresponds to Poisson regression with a relaxation of the requirement that the variance equals the mean. (See Chapter 10 of McCullach and Nelder [20].) Census blocks with populations less than four hundred were not used in the analysis, and some blocks with outlying observations were also removed.

One advantage of these models over the cluster analysis procedures discussed in Section 1.2 is that these models yield quantitative risk and rate estimates in addition to finding

clusters. As pointed out by Gangnon and Clayton (1999), modeling instead of looking for clusters also has the benefit of forcing the analyst to go through the meaningful exercise of explicitly formulating, debugging and testing the model.

In both cases, the random variables representing the incident leukemia rates are assumed to be independent across block groups.

The fitted models are below:

STF1A Model:

$$\begin{aligned} \log E[Cases_i/Pop_i] = & \beta_0 + \beta_1 Ageover60_i + \beta_2 Near.Monarch_i \\ & + \beta_3 Near.GE_i + \beta_4 Near.Solvent_i + \beta_5 Near.SmithCorona \\ & + \beta_6 Near.Morse \end{aligned}$$

$$\begin{aligned} Var[Cases_i] &= \phi E[Cases_i] \\ i &= 1, \dots, 790 \text{ census blocks} \end{aligned}$$

STF3A Model:

$$\begin{aligned} \log E[Cases_i/Pop_i] = & \beta_0 + \beta_1 AgeOver60_i + \beta_2 PublicWater_i + \beta_3 ProtectService_i \\ & + \beta_4 PrecisionRepair_i + \beta_5 Technician_i + \beta_6 TotalManufacturing_i \\ & + \beta_7 Near.Monarch_i + \beta_8 Near.GE_i \\ & + \beta_9 Near.Solvent_i + \beta_{10} Near.SmithCorona_i \end{aligned}$$

$$\begin{aligned} Var[Cases_i] &= \phi E[Cases_i] \\ i &= 1, \dots, 608 \text{ census blocks} \end{aligned}$$

2.4 Analysis

The models presented above yield two pieces of information: coefficient estimates and analysis of deviance tables. The coefficient estimates are in Table 3 and the analyses of deviance are in Table 4.

Table 3: Fitted Model Coefficient Estimates

Variable	STF1A Coeff	t-stat	p-value*	STF3A Coeff	t-stat	p-value*
AgeOver60	4.1	7.6	0.00	2.6	3.71	0.00
PublicWater				0.5	1.51	0.07
ProtectService				8.0	1.21	0.11
PrecisionRepair				-10.6	-3.91	0.00
Technicians				-8.6	-1.68	0.05
TotalManufacturing				3.1	2.18	0.01
Near.Monarch	1.0	1.39	0.08	0.7	1.00	0.16
Near.GE	0.2	1.32	0.09	0.2	0.89	0.19
Near.Solvent	0.8	0.36	0.36	1.0	0.50	0.31
Near.SmithCorona	1.9	4.29	0.00	1.6	3.44	0.00
Near.Morse	-0.1	-0.38	0.35			

*These p-values are not adjusted for the variable selection procedure.

In generalized linear models, a variable's contribution to deviance divided by total deviance with no variables in the model (null deviance) can be considered to be the percentage of fit attributable to that variable. When the design is not orthogonal or the data are not normally distributed, the deviance depends on the order in which the variables were entered into the model. To give a sense of the contribution of fit for each variable in our model, we list two deviances for each, one when it is included in the model first and one when it is put in last. Note that these values *do not* necessarily bound all the possible values of deviance for all possible orders of variable addition.

Note that for both models, the deviance attributable to distance to the TCE sites is a very small percentage of the total deviance. While there are tests of significance based on deviance, we base our tests on the t-statistics since we have such a large sample size.

Table 4: Analysis of Deviance

Variable	STF1A First	Last	STF3A First	Last
AgeOver60	0.0474	0.0429	0.0279	0.0107
PublicWater			0.0167	0.0018
ProtectService			0.0034	0.0011
PrecisionRepair			0.0203	0.0126
Technician			0.0008	0.0023
TotalManufacturing			0.0006	0.0037
Near.Monarch	0.0039	0.0015	0.0040	0.0007
Near.GE	0.0026	0.0012	0.0012	0.0006
Near.Solvent	0.0000	0.0001	0.0000	0.0002
Near.SmithCorona	0.0100	0.0111	0.0100	0.0074
Near.Morse	0.0005	0.0001		
Null Deviance	0.5781		0.4274	

2.5 Interpretation

Due to missing data, non-orthogonal design, and use of all subsets regression, interpreting the results requires extra care. Note that significant associations between sites and disease do not establish causality. See Section 2.8 for more discussion of this issue.

Consider the results for STF1A first. At a nominal 10% significance level, the variables age, near.Monarch, near.GE and near.SmithCorona are all associated with increased incident leukemia rates. Age is widely believed to be a risk factor for leukemia. The sites identified as statistically significant are consistent with the sites identified by age unadjusted studies cited in Sections 1.2 and 1.3. However, in Waller and McMaster's 1997 study, age standardization increased the estimated relationship between the site and the case count. Our age adjustment decreased the estimated relationship (see Table 5). In Table 4, note that the deviance associated with the variable Near.Monarch varies greatly depending on whether or not other covariates are used. This suggests that Near.Monarch's relation to the response is strongly influenced by the other variables in this analysis.

Consider the STF3A analysis next. Age remain significant. There is a suggestion of increased risks associated with drinking public water and working in “manufacturing”. A high percentage of the population working as technicians or in a jobs classified as precision repair are associated with lower rates. We do not have an explanation for this. In a departure from previous analyses, Table 3 shows that the only site variable associated with increased leukemia rates after controlling for other covariates is Near.SmithCorona.

Tables 3, 4, and 5, and correlation calculations, suggest that the previous results about the G.E. and Monarch sites were confounded by age and occupation. More specifically, the percentage of residents over 60 years old, the percentage working in jobs classified as precision repair and manufacturing, and drinking public water are related to both the site proximity and elevated leukemia counts. Of those variables, only removing the predictor PublicWater from the analysis had little effect on the estimates in Table 5. Hence we conclude the clustering that was attributed to site proximity might actually be due to clustering of residents in certain occupations and population age.

A natural question is whether the occupations are intervening variables between Monarch and G.E. and the elevated leukemia rates. While they probably are to a certain extent, the extent is probably limited since the employees of those two companies made up a small part of the total manufacturing employment in their respective areas.

Table 5: Summary of TCE Site Parameter Estimates

Covariate	Base Case Est / t value	STF1A Est / t value	STF3A Est / t value
Near.Monarch	1.70 / 2.42	0.98 / 1.39	0.72 / 1.00
Near.GE	0.38 / 2.07	0.24 / 1.32	0.18 / 0.89
Near.Solvent	0.08 / 0.03	0.78 / 0.36	1.03 / 0.50
Near.SmithCorona	1.89 / 3.96	1.94 / 4.29	1.60 / 3.44
Near.Morse	-0.16 / -0.50	-0.11 / -0.38	

Results differ between the STF1A and STF3A models. It is probable that some of the change in the results from STF1A to STF3A is due to a lack of power. More specifically, since STF3A has both more covariates and fewer observations, we would expect p-values to increase. The difference in results for variable near.GE may fall into this category. On the other hand, while near.Monarch’s p-value also increases, that change is accompanied by a large change in its coefficient. That suggests that the effects of being near Monarch is confounded with other variables in STF3A. It is worth pointing out that only the Morse site had more than 5% of the observations within 20 kilometers missing.

2.6 Statistical Significance and Public Health Significance

One measure of public health significance is the rate ratio:

$$\frac{\text{Incidence Rate for Exposed}}{\text{Incidence Rate for Unexposed}}$$

[14]. Under the Poisson model, the parameter estimates estimate the log of the associated rates so that the rate ratios are readily estimated. Table 6 summarizes the rate ratios associated with each variable for the STF1A and STF3A models:

Table 6: Estimated Rate Ratios

Variable	STF1A Rate Ratio	STF3A Rate Ratio
AgeOver60***	60.3	13.5
PublicWater		1.6
ProtectService		2981.0
PrecisionRepair**		2.5e5
Technicians**		1.8e3
TotalManufacturing**		22.2
Near.Monarch*	2.7	2.0
Near.GE*	1.2	1.2
Near.Solvent	2.2	2.7
Near.SmithCorona***	6.7	5.0
Near.Morse	0.9	

*Nominally significant at 10% level in STF1A analysis only.

**Nominally significant at 10% level in STF3A analysis only.

***Nominally significant at 10% level in both STF1A and STF3A.

From this table it can be seen that of the nominally significant variables, ‘AgeOver60’ and ‘TotalManufacturing’ seem to have the most public health significance. The point estimates of the rate ratios associated with the TCE sites are much smaller. Confidence intervals around these point estimates are consistent with this interpretation.

2.7 Other approaches

Other models besides generalized linear models are feasible for these data. We outline some of these below.

Cases whose addresses were unknown were allocated proportionally across the block groups in areas of likely residence. This suggests that an appropriate model should take into account the multiple resolutions at which the data were collected. Future theoretical work will make this idea more specific.

An aspect of these data which we did not consider is that even after adjusting for the covariates, the number of cases per census block group may be spatially correlated. A paper by Ghosh, et al (1999) suggests a way to address that aspect of the data using hierarchical Bayesian models (Besag et al. 1991, 1995).

Another way to relax the model assumptions would be to consider Generalized Additive Models (Hastie and Tibshirani (1990) [11]). An interesting future project would be to combine this approach with the hierarchical Bayesian approach cited above.

Finally, another fairly simple modeling approach would have been to conduct a cluster analysis on the residuals from the two regression models without the inverse distances to the TCE sites. We tried this by applying a version of the uniformly most powerful test designed by Waller (1992), modified to account for the possible negative values of the residuals. No evidence of significant clustering around any of the eleven foci was discovered.

In addition to different modeling approaches, it would be useful to investigate additional data sources as well. Newer case and covariate data would be more relevant. Also, the

exposure variables could be improved. While TCE storage sites are one source of industrial exposure into the environment, we suspect that there are many others also. The Cornell University Geospatial Information Repository contains Graphical Information System (GIS) overlays containing spatial information about annually reported industrial chemical releases in New York State. Further exploration of how GIS capabilities could be used in a study like this also would be interesting.

2.8 Weaknesses of the analysis

In addition to the statistical concerns discussed in the previous section, the primary weakness with this analysis lies in the fact that it is an ecological study. Since the cases and covariates are aggregated to the block group (or tract) level, we do not know if the covariates actually apply to the cases. Studies such as these certainly can suggest a causal relationship, but they provide much weaker evidence than a cohort or case control study for instance.

A second factor which makes it difficult to establish causality for this study is that case incidence and site exposures are measured concurrently. Better data would consider the induction period of leukemia and measure exposure before incidence.

3 Conclusion

3.1 Summary of results

Our analysis improves previous analyses of these data by taking additional relevant covariates (age, occupation, and water source) into account. While previous authors stated the desirability of adjusting for covariates only Waller and McMaster (1997) added the available census data.

After controlling for those covariates, we found a relationship between being near the Smith Corona site and an elevated leukemia rate in 1978-1982. While previous analyses had found relationships between elevated rates and other sites, our analysis suggests that those results were confounded by occupation, industry, and age.

3.2 Current status of the sites

Monarch Chemical, GE Auburn, and Smith Corona were the three sites that proved most interesting in our analyses. Monarch Chemical, which is currently on the national Comprehensive Environmental, Response, Compensation, and Liability (CERLIS) Hazardous Waste Site list, has participated in clean-up efforts. In 1982 two wells in the Vestal Water Supply system were found to be contaminated with high levels of TCE; the wells were closed until a treatment system was constructed, and were reopened in 1988. New York State took legal actions against Monarch Chemical and other potentially responsible parties, and an agreement was signed in 1985. As part of the signed agreement, the potentially responsible parties paid to have 42 tons of contaminated soil removed. Levels of contaminants in untreated ground-water have since declined to levels approaching drinking water standards, and the site remediation was considered complete in September, 1998 [26].

The site at GE Auburn is listed in the national No Further Remedial Action Planned (NFRAP) site list. The only actions listed for this site were an Initial Discovery in June, 1981 and a Preliminary Assessment in June, 1987. The Smith Corona site in Cortland is not listed on any national waste site lists; the authors are not aware of what cleanup actions, if any, have been taken.

3.3 Environmental Epidemiological Studies and Public Policy

As concern over cancer rises in our society, it becomes ever more important to address the societal impact of studies of disease clusters. Both the policy and legal impacts can directly affect individuals.

For example, the small town of Woburn Massachusetts has become the focus of much publicity beginning in 1982, when families of children with leukemia filed suit against two local companies for medical damages that they alleged resulted from the contamination of the local water wells by the companies [4] [18]. The case is now the subject of a book, *A Civil Action* [10], and a major motion picture of the same name. The Woburn case, and others like it (civil mass tort cases for environmental contamination and the damage allegedly caused to individuals as a result) rely heavily upon statistical evidence. The courts will be increasingly expected to instruct juries regarding the use of statistical evidence.

There are also political and policy ramifications. Agencies, such as environmental protection agencies, have individually tailored responsibilities and overriding management goals. However, they must also be responsive to the concerns of the citizenry and of local, state, and national political decision-makers in order to survive politically and establish a support base within the community[13]. These requirements can cause complications when dealing with cancer cluster studies for reasons that will be outlined below.

Cancer is likely to be associated with increased perceptions of risk for several reasons:

- children may be directly affected;
- individuals perceive that the disease cannot be avoided or controlled;
- it is associated with feelings of dread;
- the prevalent perception is that individuals are exposed to carcinogenic compounds without their knowledge or consent.

As cases such as Woburn become more widely known, these concerns will continue to increase as will reports of potential clusters [6]. Cancer cluster studies can place a heavy financial and time burden on agencies, especially when searching for site specific incidence. The combination of the rare disease rate and the small number of people living close to any one facility can create difficulties in even the best-designed studies. Nonsignificant findings may not lay concerns to rest in the eyes of the affected communities.

There are three reasons that "non-significant" findings may still have policy implications:

- poor public understanding of the statistical and scientific issues
- differing views of "acceptable" risk among the affected communities and the policy-making agencies

- disproportionate distribution of potential hazards in low-income or otherwise disadvantaged communities.

Agencies and individuals involved in communicating study results to the public must be aware of these problems and the potential for the findings to become part of our legal, political, and policy environments.

Whenever possible, a brief synopsis of results should be available to the public in easy-to-understand language. Increased focus on methods of risk communication specifically tailored to cancer cluster incidence may be beneficial for improved agency communication with the public before, during, and after cancer cluster studies. Agencies may be able to use these methods to effectively communicate with the public about the *need* for a study in the first place, which may mitigate some of the strain placed on agency budgets by the increased demand for cluster studies.

References

- [1] Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration with two applications to spatial statistics. *Ann. Inst. Statist. Math.* 43, 1-59.
- [2] Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science*. 10, 3-66.
- [3] Besag, J., and Newell, J. (1991). The Detection of Clusters in Rare Diseases. *J. R. Statist. Soc. A* 154, Part 1, 143-155.
- [4] Fienberg S., and Kaye, D. H. (1991). Legal and Statistical Aspects of some Mysterious Clusters. *J. R. Statist. Soc. A* 154, Part 1, 61-74.
- [5] Gangnon, R. E. and Clayton, M. K. (1998) Bayesian Spatial Disease Clustering: An Application. Technical Report #132, Department of Biostatistics, University of Wisconsin-Madison
- [6] Gawande, A. (1999). The Cancer-Cluster Myth. *The New Yorker*. 9 Feb, 34-37.
- [7] Gerba, C. P. (1996). Risk Assessment. Pepper, I.L., C.P. Gerba, and Brusseau, M.L. (eds). *Pollution Science*. Academic Press.
- [8] Ghosh, M., Natarajan, K., Waller, L., and Kim, D. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*. 75, 305-318.
- [9] Gordis, L. (1996). *Epidemiology*. W. B. Saunders Company, Philadelphia.
- [10] Harr, J. (1995). *A Civil Action*. Random House, NY.
- [11] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

- [12] Iwano, E. (1989). A Comparison of Cluster Detection Procedures. M.S. Thesis. Department of Operations Research and Industrial Engineering, Cornell University.
- [13] Jones, Charles O. (1984). *it/ An Introduction to the Study of Public Policy*. 3rd Ed. Harcourt Brace, Fort Worth.
- [14] Kelsey, J., Whittemore, A., Evans, A., and Thompson, D. (1996) *Methods in Observational Epidemiology*. Oxford University Press, New York.
- [15] Kimbrough, R., Mitchell, F., and Houk, V. (1985), Trichloroethylene: an update. *Journal of Toxicology and Environmental Health*. **15**, 369-383.
- [16] Kleinbaum, D., Kupper, L., and Morgenstern, H. (1982). *Epidemiologic Research*. Van Nostrand Reinhold, New York.
- [17] Kulldorff, M., and Nagarwalla, N. (1995). Spatial Disease Clusters: Detection and Inference. *Statistics in Medicine*. **14**, 799-810.
- [18] Lagakos, S. W., Wessen, B. J., and Zelen, M. (1986) An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). *Journal of the American Statistical Association*. **82**, p583-596.
- [19] Lange, N., Ryan, L., Billard, L., Brillinger, D., Conquest, L., and Greenhouse, J. (eds). *Case Studies in Biometry*. Wiley, New York.
- [20] McCullach, P. and Nelder, J.A. (1989) *Generalized Linear Models, Second Edition*. Chapman and Hall, London.
- [21] Openshaw, S., Craft, A. W., Charlton, M., and Birch, J. M. (1988) Investigation of leukemia clusters by use of a geographical analysis machine. *Lancet*. 272-273.
- [22] Rogerson, P. (1999) The Detection of Clusters Using a Spatial Version of the Chi-Square Goodness-of-Fit Statistic. *Geographical Analysis*. **31**, 1, 130-147.
- [23] Rothman, K. and Greenland, S. (1998) *Modern Epidemiology*. Lippincott - Raven.
- [24] Stone, R. Investigations of excess environmental risks around putative sources: statistical problems and proposed tests. *Statistics in Medicine*. **7**, 649-660.
- [25] Turnbull, B., Iwano, E., Burnett, W., Howe, H., and Clark, L. (1990). Monitoring for Clusters of Disease: Application to Leukemia Incidence in Upstate New York. *Amer. J. Epidemiol.* **132**, 4, S136-S143.
- [26] United States Environmental Protection Agency, Superfund National Priority Site Fact Sheet, Vestal Water Supply Well 42. www.epa.gov/r02earth/superfund/sitesum/0202152c.htm
- [27] United States Environmental Protection Agency, Superfund Archive Sites: No Further Remedial Action Planned (NFRAP), General Electric / Auburn Plant. www.epa.gov/oerrpage/superfund/sites/arcsites/reg02/a0201449.htm

- [28] Waller, L. (1992). Ph.D. Thesis. Department of Operations Research and Industrial Engineering, Cornell Univerisy.
- [29] Waller, L., Turnbull, B., Clark, L., and Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence in TCE-contaminated dumpsites in upstate New York. *Environmetrics*. **3**, 281-300.
- [30] Waller, L. (1996) Statistical Power and Design of Focused Clustering Studies. *Statistics in Medicine*. **15**, 765-782.
- [31] Waller, L., and McMaster, R. (1997) Incorporating indirect standardization in tests for disease clustering in a GIS environment. *Geographical Systems*, **44**, 4, 327-342.
- [32] Waller, L. and Turnbull, B. (1993) The Effects of Scale on Tests for Disease Clustering. *Statistics in Medicine*. **12**, 1869-1884.
- [33] Whittemore, A., Friend, N., Brown Jr., B., and Holly, E. (1987). A test to detect clusters of disease. *Biometrika* **74**, 3, 631-635.